2-12-02 #3



SUGHRUE, MION, ZINN, MACPEAK & SEAS, PLLC

2100 PENNSYLVANIA AVENUE, N.W. WASHINGTON, DC 20037-3213 TELEPHONE (202) 293-7060 FACSIMILE (202) 293-7860 www.sughrue.com

March 14, 2001



J. Frank Osha, Esq Direct Dial (202) 663-7915 Email: fosha@sughrue.com

BOX PATENT APPLICATION Commissioner for Patents Washington, D.C. 20231

Re:

Application of Kenji YAMANISHI and Hang LI

OUESTIONNAIRE ANALYSIS SYSTEM

Our Ref. Q63084

Dear Sir:

Attached hereto is the application identified above including 30 sheets of the specification, including the claims and abstract, 10 sheets of drawings, the executed Assignment and PTO 1595 form, and the executed Declaration and Power of Attorney. Also enclosed is an Information Disclosure Statement with form PTO-1449 and references.

The Government filing fee is calculated as follows:

Total claims	13 - 20	= x	\$18.00 =	
Independent claims	4 - 3	= 1 x	\$80.00 =	\$80.00
Base Fee				\$710.00
TOTAL FILING FI	E I E			\$790.00
TOTAL FILING FEE Recordation of Assignment				\$40.00
TOTAL FEE				\$830.00

Checks for the statutory filing fee of \$790.00 and Assignment recordation fee of \$40.00 are attached. You are also directed and authorized to charge or credit any difference or overpayment to Deposit Account No. 19-4880. The Commissioner is hereby authorized to charge any fees under 37 C.F.R. §§ 1.16 and 1.17 and any petitions for extension of time under 37 C.F.R. § 1.136 which may be required during the entire pendency of the application to Deposit Account No. 19-4880. A duplicate copy of this transmittal letter is attached.

Priority is claimed from March 15, 2000 based on Japanese Application No. 2000-071657. The priority document is enclosed herewith.

> Respectfully submitted, SUGHRUE, MION, ZINN, MACPEAK & SEAS, PLLC

Attorneys for Applicant

Registration No. 24,625

Yamanishi et al Filed 3/14/01 Q 63084

日本国特許庁 PATENT OFFICE

PATENT OFFICE
JAPANESE GOVERNMENT



別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 Date of Application:

1

2000年 3月15日

出 願 番 号 Application Number:

特願2000-071657

出 願 人 Applicant (s):

日本電気株式会社

2000年12月 1日

特許庁長官 Commissioner, Patent Office





特2000-07165

【書類名】

特許願

【整理番号】

33509710

【あて先】

特許庁長官殿

【国際特許分類】

G06F 17/00

【発明者】

【住所又は居所】

東京都港区芝五丁目7番1号 日本電気株式会社内

【氏名】

山西 健司

【発明者】

【住所又は居所】

東京都港区芝五丁目7番1号 日本電気株式会社内

【氏名】

李 航

【特許出願人】

【識別番号】

000004237

【氏名又は名称】 日本電気株式会社

【代理人】

【識別番号】

100088890

【弁理士】

【氏名又は名称】

河原 純一

【手数料の表示】

【予納台帳番号】

009690

【納付金額】

21,000円

【提出物件の目録】

【物件名】

明細書 1

【物件名】

図面 1

【物件名】

要約書 1

【包括委任状番号】 9001717

【プルーフの要否】

要

【書類名】 明細書

【発明の名称】アンケート回答分析システム

【特許請求の範囲】

【請求項1】自然言語による自由回答記述を含むアンケート回答文を入力する手段と、

前記アンケート回答文を送信するネットワークと、

前記送信されたアンケート回答文を蓄積するデータベースと、

該データベースから前記アンケート回答文を読み出して、前記アンケート回答文 を分類するルールを学習するテキスト分類エンジンと

を備えることを特徴とするアンケート回答分析システム。

【請求項2】自然言語による自由回答記述を含むアンケート回答文を入力する手段と、

前記アンケート回答文を蓄積するデータベースと、

該データベースから前記アンケート回答文を読み出して、前記アンケート回答文 を分類するルールを学習するテキスト分類エンジンと

を備えることを特徴とするアンケート回答分析システム。

【請求項3】自然言語による自由回答記述を含むアンケート回答文を入力する手段と、

前記アンケート回答文を送信するネットワークと、

前記送信されたアンケート回答文を蓄積するデータベースと、

該データベースから前記アンケート回答文を読み出して、前記アンケート回答文 を分類するルールを学習するテキスト分類エンジンと、

該ルールを要求者からの要求に応じて前記ネットワークを通じて配信する手段と

を備えることを特徴とするアンケート回答分析システム。

【請求項4】前記テキスト分類エンジンが、前記データベースに蓄積されたアンケート回答文の全文に対して形態素解析を行う形態素解析手段と、カテゴリとテキストとを指定させるカテゴリ・テキスト指定手段と、前記データベースから読み込んだ複数のアンケート回答文に対して属性選択を行う属性選択手段と、前記

属性選択手段により属性選択された単語を基にテキストとカテゴリとの対応を表すルールを学習するルール学習手段と、前記ルール学習手段により学習されたルールを出力するルール出力手段とを含む請求項1,請求項2,または請求項3記載のアンケート回答分析システム。

【請求項 5】前記属性選択手段が、テキスト中に出現する単語の各々について、 該単語の出現を考慮しない場合のテキスト集合の確率的コンプレキシティと考慮 する場合のテキスト集合の確率的コンプレキシティとの差 Δ S C (ω) を計算す る手順と、該差 Δ S C (ω) がしきい値 τ より大きければ属性として選択する手 順とを実行する請求項 4 記載のアンケート回答分析システム。

【請求項6】前記ルール学習手段が、テキストの集合を、(d_1 , c_1),(d_2 , c_2),…,(d_m , c_m)[ただし、各 d_i は多値の離散ベクトル d_i = (ω_{i1} , ω_{i2} , …, ω_{in})(i=1, …, m)、 ω_{ij} は属性選択で得られた単語 ω_{j} (j=1, …, n)がi番目のテキストに現れれば1、そうでなければ0の値をとる、 c_i はi番日のテキストに対応するカテゴリの値(ラベル)を表し、各 c_i は所与のカテゴリに属するならば1、そうでなければ0の値をとる、mはテキストの数]の表現に置き換えて成形する手順と、拡張型確率的コンプレキシティ最小原理または確率的コンプレキシティ最小化原理などの情報量規準を用いてif-then-else型のルールを選択して確率的決定リストに順番に追加していく成長処理を行う手順と、前記確率的決定リストの最後からルールを1つずつ取り除き拡張型確率的コンプレキシティ最小原理の観点からみてこれを取り除かないほうがよいところまで続ける刈り込み処理を行う手順とを実行する請求項4記載のアンケート回答分析システム。

【請求項7】コンピュータを、前記データベースに蓄積されたアンケート回答文の全文に対して形態素解析を行う形態素解析手段,前記テキスト分類エンジンに対してカテゴリとテキストとを指定させるカテゴリ・テキスト指定手段,前記データベースから読み込んだ複数のアンケート回答文に対して属性選択を行う属性選択手段,前記属性選択手段により属性選択された単語を基にテキストとカテゴリとの対応を表すルールを学習するルール学習手段,および前記ルール学習手段により学習されたルールを出力するルール出力手段として機能させるためのプロ

グラムを記録した記録媒体。

【発明の詳細な説明】

[0001]

【発明の属する技術分野】

本発明はアンケート回答分析システムに関し、特にテキスト自動分類技術、自然言語処理技術、およびネットワーク利用技術を使用するアンケート回答分析システムに関する。

[0002]

【従来の技術】

インターネット等のネットワークを通じて得られる、自然言語による自由回答記述を含むアンケート回答文から全体的な特徴や傾向を導き出す作業は、従来の多くは人手で行われてきた。 株式会社電通が開発したDEーFACTO(パンフレット)や富士通株式会社のKeyword Associator(渡部勇:発散的思考支援システム「Keyword Associator」第二版,計測自動制御学会第15回システム工学部会研究会資料,1994年7月),HIPS等のテキストマイニングツール(渡辺,三末,新田,杉山: ハイブリッド発想支援システム「HIPS」,計測自動制御学会第15回システム工学部会研究会資料,1995年1月)は、テキスト情報から言葉の関係性を導き出すのに用いられた。しかし、これらのツールは、ルールという形式でアンケート回答文の特徴を表現するものではなかった。

[0003]

また、自然言語による自由回答記述を含むアンケート回答文をインターネット等のネットワークを通じて自動的に収集して分析し、場合によっては要求者に分析結果を配信するようなシステムやサービスはこれまで存在しなかった。たとえば、特開平11-066036号公報,特開平11-143856号公報等には、ネットワークを通してアンケートを実施し、回答をデータベースに蓄積する技術が開示されているが、ルールという形式でアンケート回答文の特徴を抽出することまでは行われていない。

[0004]

【発明が解決しようとする課題】

上述した従来の人手によるアンケート回答分析では、アンケート回答文の数が大量になると、非効率で手に負えないという問題点があった。

[0005]

また、DE-FACTOやHIPS等のテキストマイニングツールでは、ルールという形式でアンケート回答文の特徴を抽出できないので、コンパクトで分かりやすい知識の提示という点で十分ではないという問題点があった。

[0006]

さらに、検索等で用いられているテキスト分類ツールは、アンケート回答分析に も有力なツールであるが、自然言語による自由回答記述を含むアンケート回答文 の分析に用いられている前例はなかった。

[0007]

本発明の第1の目的は、テキスト分類エンジンを用いることによって、自然言語による自由回答記述を含むアンケート回答文からコンパクトで分かりやすいルール形式の知識を自動的に提示することを可能にするアンケート回答分析システムを提供することにある。

[0008]

また、本発明の第2の目的は、上記に加えて、ネットワークを通じて集めたアンケート回答文からルール形式の知識を自動的に抽出し、要求者に分析結果を配信するアンケート回答分析システムを提供することにある。

[0009]

【課題を解決するための手段】

本発明のアンケート回答分析システムは、自然言語による自由回答記述を含むアンケート回答文を入力する手段と、前記アンケート回答文を送信するネットワークと、前記送信されたアンケート回答文を蓄積するデータベースと、該データベースから前記アンケート回答文を読み出して、前記アンケート回答文を分類するルールを学習するテキスト分類エンジンとを備えることを特徴とする。

[0010]

また、本発明のアンケート回答分析システムは、自然言語による自由回答記述を

含むアンケート回答文を入力する手段と、前記アンケート回答文を蓄積するデータベースと、該データベースから前記アンケート回答文を読み出して、前記アンケート回答文を分類するルールを学習するテキスト分類エンジンとを備えることを特徴とする。

[0011]

さらに、本発明のアンケート回答分析システムは、自然言語による自由回答記述を含むアンケート回答文を入力する手段と、 前記アンケート回答文を送信するネットワークと、前記送信されたアンケート回答文を蓄積するデータベースと、該データベースから前記アンケート回答文を読み出して、前記アンケート回答文を分類するルールを学習するテキスト分類エンジンと、該ルールを要求者からの要求に応じて前記ネットワークを通じて配信する手段とを備えることを特徴とする。

[0012]

【発明の実施の形態】

以下、本発明の実施の形態について図面を参照して詳細に説明する。

[0013]

(1)第1の実施の形態

図1は、本発明の第1の実施の形態に係るアンケート回答分析システムの全体構成を示すブロック図である。本実施の形態に係るアンケート回答分析システムは、回答者コンピュータ111~11N(Nは正整数)と、ネットワーク12と、データベース13と、テキスト分類エンジン14とから、その主要部が構成されている。

[0014]

回答者コンピュータ111~11Nは、コンピュータ,携帯情報端末,携帯電話機等の各種のメッセージ,メール等の送信機能を持つ機器が該当し、ネットワーク12に接続されている。

[0015]

ネットワーク12には、有線、無線を問わない各種の公衆回線網、専用線網や、LAN(Local Area Network)等の各種のネットワークが含

まれる。

[0016]

データベース13は、ネットワーク12に接続されており、回答者コンピュータ 111~11Nからネットワーク12を通じて送信された、複数のアンケート回答者からのアンケート回答文を蓄積する。

[0017]

テキスト分類エンジン14は、データベース13から複数のアンケート回答文を読み出して、アンケート回答文を分類するルールを導き出し、ルールを要求者に出力する。テキスト分類エンジン14は、データベース13に蓄積されたアンケート回答文の全文に対して形態素解析を行う形態素解析手段15と、テキスト分類エンジン14に対してカテゴリとテキストとを指定させるカテゴリ・テキスト指定手段16と、データベース13から読み込んだ複数のアンケート回答文に対して属性選択を行う属性選択手段17と、属性選択手段17により属性選択された単語を基にテキストとカテゴリとの対応を表すルールを学習するルール学習手段18と、ルールを出力するルール出力手段19とを含んで構成されている。

[0018]

なお、テキスト分類エンジン14は、カテゴリとテキストとの間の対応関係を分類則として学習するエンジンであり、たとえば、Li and Yamanishiにより提案されているエンジン(h. Li and k. Yamanishi: Text Classification Using ESC—based Stochastic Decision Lists, Proccedings of 1999 International Conference on Information & Knowledge Management, pp:122—130, 1999)を用いることができる。このテキスト分類エンジン14は、基本的に特許第2581196号に開示された「決定リストの生成方法及び装置」の方式を利用したものである。

[0019]

図2は、データベース13に格納されるアンケート回答文の集合の構造を示す。 各列はアンケート項目を表し、各行は1人分のアンケート回答文を表す。

[0020]

図3を参照すると、テキスト分類エンジン14の処理は、形態素解析ステップ3 1と、テキストとカテゴリとの指定ステップ32と、属性選択ステップ33と、 ルール学習ステップ34と、ルール出力ステップ35とからなる。

[0021]

図4を参照すると、属性選択ステップ33のより詳しい処理は、ΔSC(ω)計算ステップ41と、属性選択ステップ42とからなる。

[0022]

図5を参照すると、ルール学習ステップ34のより詳しい処理は、データ成形ステップ51と、成長処理ステップ52と、刈り込み処理ステップ53とからなる

[0023]

図6は、テキスト分類エンジン14による分析結果であるルール形式の知識(確率的決定リスト)の一例を示す図である。

[0024]

図7は、テキスト分類エンジン14による分析結果であるルール形式の知識(確率的決定リスト)の他の例を示す図である。

[0025]

次に、このように構成された第1の実施の形態に係るアンケート回答分析システムの動作について説明する。

[0026]

アンケート回答者が、回答者コンピュータ111~11Nからアンケート回答文を送信すると、アンケート回答文はネットワーク12を通じてデータベース13に蓄積される。ここで、アンケート回答者の数は、N人とする。この際、アンケート回答文は、自然言語による自由回答記述を含むものであってよい。

[0027]

テキスト分類エンジン14は、まず、形態素解析手段15により、データベース 13に蓄積されたアンケート回答文の全文に対して形態素解析を行う(ステップ 31)。

[0028]

次に、テキスト分類エンジン14は、カテゴリ・テキスト指定手段16により、オペレータにアンケート回答文中のカテゴリとテキストとを指定させる(ステップ32)。ここで、カテゴリの指定とは、1つの列の回答に注目して分類するものである。例えば、図2の1列目に注目して、その回答を「A社」と「A社以外」とに分類することが、カテゴリ指定である。また、テキスト指定とは、カテゴリ指定に用いられた列を除いて、分析に用いる列を指定することである。例えば、図2の2列目を選択してテキスト指定を行う。

[0029]

続いて、テキスト分類エンジン14は、属性選択手段17により、データベース 13から読み込んだ複数のアンケート回答文に対して属性選択を行う(ステップ 33)。ここで、属性選択とは、テキストとカテゴリとの対応を表すのに重要な 単語を選択することである。

[0030]

次に、テキスト分類エンジン14は、ルール学習手段18により、属性選択された単語を基にテキストとカテゴリとの対応を表すルールを学習する(ステップ34)。例えば、上記のカテゴリ指定とテキスト指定とをしたときに、図6のようなルールが得られる。

[0031]

図6のルールは、先ず最初の行を読んで、テキストの中で「使いやすい」という単語が現れていれば、そのような回答をした人の92.0%がハイテク企業としてA社を想起している、ということを示す。もし、「使いやすい」という単語が現れていなければ、次に、「未来」と「プライベート」という単語が同時に現れているかどうかをチェックし、現れていれば、そのような回答をした人の87.2%がハイテク企業としてA社を想起している、ということを示す。以下同様に、ifーthenーelse型のルールに従って、上から下へと条件文を読んで行く。このようなルールは、ハイテク企業とハイテク感との間の関係を分かりやすくコンパクトに示すものである。

[0032]

また、別の会社であるB社をとりあげて、「B社」と「B社以外」とにカテゴリ 指定したときに、同様な手順で図7のルールが得られたとする。

[0033]

図7のB社のルールを図6のA社のルールと比較すると、A社をハイテク企業と 想起する人のハイテク感は、使い勝手や嗜好品感覚を重視するのに対し、B社を ハイテク企業と想起する人のハイテク感は効率性を重視していることがわかる。 このように、ルールの比較により、アンケート回答分析が容易になる。

[0034]

最後に、テキスト分類エンジン14は、ルール出力手段19により、分析結果の ルール形式の知識を要求者に出力する(ステップ35)。

[0035]

ここで、ルール形式の知識の例として確率的決定リストをとり上げ、これを生成するための属性選択ステップ33およびルール学習ステップ34についてより詳しく説明する。

[0036]

確率的決定リストとは、図 6 に示すような i f - t h e n 型の確率的ルールの順序付きリストのことである。各確率的ルールは、" c = 1 \leftarrow t (確率 p)"の形をとる。ここで、c = 1 は分類の決定、t は分類の条件、(確率 p)は確率を表す。

[0037]

まず、属性選択ステップ33について詳しく述べる。

[0038]

属性選択とは、与えられたカテゴリ(例えば、'A社'と'A社でない')に対して、そのカテゴリと深く関係する単語を集めることである。具体的には、図4に示すように、ステップ41で、テキストに出現する単語 ω の各々について、その単語 ω の出現を考慮しない場合のテキスト集合の確率的コンプレキシティ(SCとかく)と考慮する場合のSCとの差 Δ SC(ω)を計算し、ステップ42で、その差 Δ SC(ω)が、与えられたしきい値 τ より大きければ単語 ω を属性として選択する。

[0039]

以下、SCの具体的な計算の仕方を述べる。入力されたアンケート回答文中のテキストの集合を

$$(d_1, c_1), (d_2, c_2), ..., (d_m, c_m)$$

と表す。ここで、 d_i は i 番目のテキストを表し、i 番目のテキストに現れる単語の列として表現される。 c_i は i 番目のテキストに対応するカテゴリの値(ラベル)を表し、各 c_i は所与のカテゴリ('A社')に属するならば i 1、そうでなければ('A社でない') i 0 の値をとる。i mはテキストの数である。

[0040]

さらに、ラベル列を $\mathbf{c}^{\ m}=\mathbf{c}_{\ 1}$, …, $\mathbf{c}_{\ m}$ 、テキスト列を $\mathbf{d}^{\ m}=\mathbf{d}_{\ 1}$, …, $\mathbf{d}_{\ m}$ と書く。ラベル列 $\mathbf{c}^{\ m}$ の \mathbf{S} \mathbf{C} を、数 $\mathbf{1}$ のように計算する。

[0041]

【数1】

$$SC(c^m) = mH\left(\frac{m^+}{m}\right) + \frac{1}{2}\log\frac{m}{2\pi} + \log\pi$$

[0042]

ここでは、 m^+ はラベル列 c^m 中で値が 1 であるラベルの数を表す。また、 1 o g は自然対数を表し、数 2 とする。

[0043]

【数2】

$$H(z) = -z \log z - (1-z) \log(1-z)$$

[0044]

たとえば、J. Rissanen, Fisher information and stochastic complexity. IEEE Trans.
on Infomation Theory, 42(1):40-47(1996)に述べられているように、SC(c^m)は、ラベル列c^mを与えられたモ

デル(ここではベルヌイモデル)を用いて記述するのに必要な最短の記述長という意味をもつ。また、

【外1】

c^m_o

は、対応するテキスト d $_i$ の中に単語 ω が現れるラベル c $_i$ からなるラベル列であるとする。ここに、 m $_\omega$ は、

【外2】

 $c^{m_{\omega}}$

におけるラベルの数であるとする。すると、

【外3】

 $c^{m_{\omega}}$

のSCの値を、数3のように計算することができる。

[0045]

【数3】

$$SC(c^{m_{\omega}}) = m_{\omega}H\left(\frac{m_{\omega}^{+}}{m_{\omega}}\right) + \frac{1}{2}\log\frac{m_{\omega}}{2\pi} + \log\pi$$

[0046]

ここでは、 m_{ω}^{+} は、

【外4】

 $c^{m_{\omega}}$

における値が1であるラベルの数を表す。一方、

【外5】

 $c^{m_{\neg \omega}}$

は、対応するテキスト d $_i$ に単語 $_{\omega}$ が現れないラベル c $_i$ からなるラベル列であるとする。 m $_{\omega}$ は、

【外6】

におけるラベルの数である。

[0047]

【外7】

$$c^{m_{\neg \omega}}$$

のSCの値を、数4のように計算することができる。

[0048]

【数4】

$$SC(c^{m_{\neg\omega}}) = m_{\neg\omega}H\left(\frac{m_{\neg\omega}^{+}}{m_{\neg\omega}}\right) + \frac{1}{2}\log\frac{m_{\neg\omega}}{2\pi} + \log\pi$$

[0049]

単語 ω の出現を考慮しない場合のSCと考慮する場合のSCとの差 Δ SC (ω)は、数5のように計算される。

[0050]

【数5】

$$\Delta SC(\omega) = \frac{1}{m} \left(SC(c^{m}) - \left(SC(c^{m_{\omega}}) + SC(c^{m_{-\omega}}) \right) \right)$$

$$= \left[H\left(\frac{m^{+}}{m}\right) - \frac{m_{\omega}}{m} H\left(\frac{m_{\omega}^{+}}{m_{\omega}}\right) - \frac{m_{-\omega}}{m} H\left(\frac{m_{-\omega}^{+}}{m_{-\omega}}\right) \right]$$

$$- \left[\frac{1}{2m} \log \frac{m_{\omega} m_{-\omega} \pi}{2m} \right]$$

[0051]

差 Δ S C (ω) の大きい単語 ω は、与えられたカテゴリによく現れる、あるいはほとんど現れない単語である。それらの単語がそのカテゴリと深く関係するとみることができる。そこで、 τ を与えられた閾値とし、 Δ S C (ω) $> \tau$ なる単語 ω を属性として選択する。

[0052]

次に、ルール学習ステップ34について詳しく述べる。

[0053]

いま、属性選択された単語がn 個あるとし、 ω_1 , …, ω_n とする。ステップ 5 1 で、まず、入力されたテキストの集合を、以下の表現に置き換える。

 $(d_1, c_1), (d_2, c_2), \dots, (d_m, c_m)$

[0054]

ここで、各 d i は、 2 値ベクトル (一般的に多値の離散ベクトル)

 $d_{i} = (\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,n}) \quad (i = 1, \dots, m)$

を表す。ここで、 ω_{ij} は、属性選択で得られた単語 ω_{j} が i 番目のテキストに現れれば 1 、そうでなければ 0 の値をとる(j=1, …, n)。 c_{i} は i 番日のテキストに対応するカテゴリの値(ラベル)を表し、各 c_{i} は所与のカテゴリに属するならば 1 、そうでなければ 0 の値をとる。m はテキストの数である。

[0055]

ステップ52では、ifーthenーelse型のルールを選択し、確率的決定リストAに順番に追加していく。これを「成長」とよぶ。ルールの選択には、例えば、拡張型確率的コンプレキシティ(ESCとかく)最小原理を用いる。

[0056]

これは、以下のように行う。kを与えられた正の整数とする。属性選択された単語 ω を基に可能なすべてのk項(単語 ω のk個までの連言)の集合をTとする。 次に、集合Tの項tの中からテキストに一度も現れないものを取り除く。また、空の確率的決定リストAを用意する。次に、ESCの値の減少分がもっとも大き いルールを確率的決定リストAに順次追加する。

[0057]

ここで、ESCの計算は、以下のように行う。全データ集合Dを

[0058]

【数6】

$$ESC(c^{m}) = Loss(c^{m}) + \lambda \sqrt{m \log m}$$

[0059]

これは、K. Yamanishi, A decision-theoretic extension of stochastic complexity, and its applications to learning, IEEE Trans. Inform. Theory, 44, 1424-1439 (1998) の論文で提案されたオリジナルのESCの1つの近似形である。

[0060]

ここでは、 λ は正の定数を表し、L o s s (c m) はデフォルトの分類を行う際の誤りの数を表す。デフォルトの分類とは、たとえば、すべてのラベルが0 であると仮定することである。 t は集合Tの中の項であるとする。

[0061]

【外8】

 c^{m_t}

は、対応するテキスト d $_i$ において項 t が真になるラベル c $_i$ からなるラベル列 であるとする。ここに、 m $_t$ は、

【外9】

cmt

におけるラベルの数であるとする。

[0062]

特2000-071657

【外10】

Loss (c^m,)

は、項tによる分類を行う際の誤りの数であるとする。一方、

【外11】

 $c^{m_{\neg \omega}}$

は、対応するテキスト d_i において項 t が偽になるラベル c_i からなるラベル列 であるとする。ここに、 m_{-t} は、

【外12】

 $c^{m_{\neg \omega}}$

におけるラベルの数であるとする。¬tは項tの否定を表す。

[0063]

【外13】

Loss (c^m¬¹)

は、¬tによる分類を行う際の誤りの数であるとする。

[0064]

【外14】

 c^{m_t}

と

【外15】

 $c^{m_{\neg t}}$

とのESCの値を、それぞれ数7および数8のように計算することができる。

[0065]

【数7】

$$ESC(c^{m_t}) = Loss(c^{m_t}) + \lambda \sqrt{m_t \log m_t}$$

[0066]

【数8】

$$ESC(c^{m_{\neg t}}) = Loss(c^{m_{\neg t}}) + \lambda \sqrt{m_{\neg t} \log m_{\neg t}}$$

[0067]

項tによる分類を行う場合、ESCの値の減少分ΔESC(t)を、数9で計算する。

[0068]

【数9】

$$\Delta ESC(t) = ESC(c^{m}) - (ESC(c^{m_t}) + ESC(c^{m_{\neg t}}))$$

$$= \left[Loss(c^{m}) - Loos(c^{m_t}) - Loos(c^{m_{\neg t}})\right]$$

$$+ \left[\lambda(\sqrt{m\log m} - \sqrt{m_t \log m_t} - \sqrt{m_{\neg t} \log m_{\neg t}})\right]$$

[0069]

ESC最小原理では、 ΔESC (t)が最小になるように項tを選択する。そのようなt=t*が選ばれたならば、これを真とするような全データ集合Dのデータの数を

【外16】

m,

とし、そのようなデータのうち、かりに数が多い方のラベルを c=1 として、 c=1 であった数を

【外17】

m₁. †

と、 c = 0 であった数を

【外18】

 m_{i}

とする。そこで、ルール" $c = 1 \leftarrow t$ * (確率 p) "を確率的決定リストAに追加する。ここで、確率値 p はラプラス推定の方法を用いて、例えば、数 1 O により計算する。

[0070]

【数10】

$$(m_{\star}^{+} + 0.5)/(m_{\star}^{+} + 1)$$

[0071]

集合Tから項t*を除いたものを新しい集合Tとし、全データ集合Dから項t* を真にする全てのデータを除いたものを新しい全データ集合Dとして、全データ 集合Dが空になるまで、同じ操作を繰り返す。なお、上記に用いた基準ESCの 代わりに、属性選択に用いた基準SCを用いることもできる。

[0072]

次に、ステップ53では、ステップ52で得られた確率的決定リストAが学習データに過度に適合することがあるので、確率的決定リストAの最後からルールを1つずつ取り除き、ESC最小原理の観点からみてこれを取り除かないほうがよいところまで続ける。これを「刈り込み」とよぶ。

[0073]

この場合のESC最小原理の適用の仕方を、以下に述べる。まず、ラベル列 e^{m} の確率的決定リストAに対するESCの値を、数11のように、確率的決定リストAにおけるすべての項 t に対するESCの値の和として定義する。

[0074]

【数11】

$$ESC(c^{m}|A) = \sum_{t} ESC(c^{m_{t}})$$

[0075]

ただし、

【外19】

$$ESC(c^{m_t})$$

は、数7のように計算する。次に、ラベル列 c^m と確率的決定リストAとの全ES C値を、数12のように定義する。

[0076]

【数12】

$$ESC(c^{m}:A) = ESC(c^{m}|A) + \lambda' L(A)$$

$$= \sum_{t} Loss(c^{m_{t}}) + \lambda \sum_{t} \sqrt{m_{t} \log m_{t}} + \lambda' L(A)$$

[0077]

ここでは、 λ は正の定数を表す。 L (A) は、確率的決定リストAを符号化する時の符号長である。具体的には、L (A) = l o g T + l o g T (T - i + l) と計算する。ただし、T は可能な項 t の数で、i は確率的決定リストAにおけるルールの数である。

[0078]

Aは刈り込み前の確率的決定リストを、A'は刈り込み後の確率的決定リストを表すものとする。

$$ESC(c^m:A) \leq ESC(c^m:A')$$

[0079]

すなわち、

$$ESC(c^{m}|A')-ESC(c^{m}|A) \ge \lambda'(L(A)-L(A'))$$

が成り立つ限り、刈り込み処理を続け、この条件が満足されなくなったか、刈り込むべきルールが無くなった時点で得られている確率的決定リストAを出力する。ESCが全体として小さい確率的決定リストAが出力されることになる。

[0800]

第1の実施の形態に係るアンケート回答分析システムでは、ネットワーク12を 通じて収集した自然言語による自由回答記述を含むアンケート回答文から自動的 に分析結果のルールを導くことができる。

[0081]

第1の実施の形態に係るアンケート回答分析システムにおいて、テキスト分類エンジン14として、例えば、Li and Yamanishiにより提案されているエンジン(h. Li and k. Yamanishi: Text Classification Using ESC—based Stochastic Decision Lists, Proceedings of 1999 International Conference on Information & Knowledge Management, pp:122—130, 1999)を用いると、O($n^k m$)の計算量で、高速にアンケート回答文からルールを導き出すことができる。ここに、nはアンケート回答文から属性選択された単語の数、mはアンケート回答文の数、kはルール1つの条件にとられる連言に含まれる単語の最大数である。よって、効率的な自動アンケート回答分析が可能となる。また、得られたルールは特定のカテゴリに属するアンケート回答文をif—then—else型のルール形式で、コンパクトにかつ分かりやすく表現している。

[0082]

第1の実施の形態に係るアンケート回答分析システムは、例えば、以下のような ビジネスとして展開できる。企業に関するイメージ調査、特定の商品やサービス 等のアンケートを請け負って、図2のようなアンケート項目でアンケートをネッ トワーク12上で実施し、ネットワーク12を通じてオンラインで集められた自 然言語による自由回答記述を含むアンケート回答文をデータベース13に蓄え、 そこからアンケート回答文を呼び出して、テキスト分類エンジン14を使用する ことにより得られるルールを分析結果として要求者に販売する。

[0083]

(2) 第2の実施の形態

図8は、本発明の第2の実施の形態に係るアンケート回答分析システムの全体構成を示すブロック図である。本実施の形態に係るアンケート回答分析システムは、アンケート回答入力手段81と、データベース82と、テキスト分類エンジン83とから、その主要部が構成されている。

[0084]

アンケート回答入力手段81は、ネットワークを介さずにデータベース82と直接接続されている。

[0085]

データベース82は、複数のアンケート回答者からのアンケート回答文を蓄積する。

[0086]

テキスト分類エンジン83は、図1に示した第1の実施の形態に係るアンケート 回答分析システムにおけるテキスト分類エンジン14と全く同様のものである。 したがって、対応する部分には同一符号を付して、それらの詳しい説明を省略す る。

[0087]

次に、このように構成された第2の実施の形態に係るアンケート回答分析システムの動作について説明する。

[0088]

アンケート回答入力手段81は、ネットワークを介さずにデータベース82と直接接続して、自然言語による自由回答記述を含むアンケート回答文を入力する。

[0089]

データベース82は、複数のアンケート回答者からのアンケート回答文を蓄積する。

[0090]

テキスト分類エンジン83は、データベース82から複数のアンケート回答文を 読み出して、アンケート回答文を分類するルールを導き出し、分析結果のルール を要求者に出力する。なお、テキスト分類エンジン83の動作の詳細は、第1の 実施の形態に係るアンケート回答分析システムにおけるテキスト分類エンジン1 4の場合と全く同様であるので、その詳しい説明を割愛する。

[0091]

第2の実施の形態に係るアンケート回答分析システムは、例えば、以下のような ビジネスとして展開できる。企業に関するイメージ調査、特定の商品やサービス 等のアンケートを請け負って、図2のようなアンケート項目でアンケートを実施 し、自然言語による自由回答記述を含むアンケート回答文を一度に回収して、これをデータベース82で蓄え、そこからアンケート回答文を呼び出して、テキス ト分類エンジン83を使用することにより得られる分析結果を要求者に販売する

[0092]

(3) 第3の実施の形態

図9は、本発明の第3の実施の形態に係るアンケート回答分析システムの全体構成を示すブロック図である。本実施の形態に係るアンケート回答分析システムは、回答者コンピュータ911~91Nと、ネットワーク92と、データベース93と、テキスト分類エンジン94と、要求者コンピュータ95とから、その主要部が構成されている。

[0093]

回答者コンピュータ911~91Nは、コンピュータ,携帯情報端末,携帯電話機等の各種のメッセージ,メール等の送信機能を持つ機器が該当し、ネットワーク92に接続されている。

[0094]

ネットワーク92には、有線、無線を問わない各種の公衆回線網、専用線網や、 LAN等の各種のネットワークが含まれる。

[0095]

データベース93は、ネットワーク92に接続されており、回答者コンピュータ

911~91Nからネットワーク92を通じて送信された、複数のアンケート回答者からのアンケート回答文を蓄積する。

[0096]

テキスト分類エンジン94は、ルール出力手段19が分析結果のルール形式の知識をネットワーク92を通じて送信できる点を除いては、図1に示した第1の実施の形態に係るアンケート回答分析システムにおけるテキスト分類エンジン14と全く同様のものである。したがって、対応する部分には同一符号を付して、それらの詳しい説明を省略する。

[0097]

要求者コンピュータ95は、ネットワーク92を通じてテキスト分類エンジン94に分析結果のルール形式の知識を要求し、テキスト分類エンジン94から分析結果のルール形式の知識をネットワーク92を通じて受信することができる。

[0098]

次に、このように構成された第3の実施の形態に係るアンケート回答分析システムの動作について説明する。

[0099]

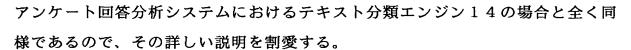
アンケート回答者は、回答者コンピュータ911~91Nからネットワーク92 を介して自然言語による自由回答記述を含むアンケート回答文を送る。なお、アンケート回答者は、N人とする。

[0100]

データベース93は、ネットワーク92に接続されており、複数のアンケート回 答者からのアンケート回答文を蓄積する。

[0101]

テキスト分類エンジン94は、データベース93から複数のアンケート回答文を 読み出して、アンケート回答文を分類するルール形式の知識を導き出す。テキスト分類エンジン94は、ネットワーク92に接続し、分析結果のルール形式の知識を要求者コンピュータ95の要求に応じてネットワーク92を通じて配信する。なお、テキスト分類エンジン94の動作の詳細は、分析結果のルール形式の知識がネットワーク92を通じて配信される点を除いて、第1の実施の形態に係る



[0102]

第3の実施の形態にアンケート回答分析システムは、例えば、以下のようなビジネスとして展開できる。企業に関するイメージ調査、特定の商品やサービス等のアンケートを請け負って、図2のようなアンケート項目でアンケートをネットワーク92上で実施し、ネットワーク92を通じてオンラインで集められた自然言語による自由回答記述を含むアンケート回答文をデータベース93で蓄え、そこからアンケート回答文を呼び出して、テキスト分類エンジン94を使用することにより得られる分析結果を、要求があれば、ネットワーク92を通じて要求者に配信サービスすることによりビジネスを行う。

[0103]

(4) 第4の実施の形態

図10は、本発明の第4の実施の形態に係るアンケート回答分析システムの全体構成を示すブロック図である。本実施の形態に係るアンケート回答分析システムは、図1に示した第1の実施の形態に係るアンケート回答分析システムにおいて、データベース13に接続されたコンピュータ101にテキスト分類エンジンプログラムを記録した記録媒体102を備えるようにしたものである。その他の構成は、第1の実施の形態に係るアンケート回答分析システムと全く同様であるので、対応する部分には同一符号を付してそれらの詳しい説明を省略する。

[0104]

このように構成された第4の実施の形態に係るアンケート回答分析システムでは、記録媒体102からコンピュータ101にテキスト分類エンジンプログラムが読み込まれ、形態素解析手段15,カテゴリ・テキスト指定手段16,属性選択手段17,ルール学習手段18,およびルール出力手段19を含むテキスト分類エンジン14としてコンピュータ101の動作を制御する。コンピュータ101上でのテキスト分類エンジン14の動作の詳細は、第1の実施の形態に係るアンケート回答分析システムの場合と全く同様になるので、その詳しい説明を割愛する。



(5) 第5の実施の形態

図11は、本発明の第5の実施の形態に係るアンケート回答分析システムの全体構成を示すブロック図である。本実施の形態に係るアンケート回答分析システムは、図8に示した第2の実施の形態に係るアンケート回答分析システムにおいて、データベース82に接続されたコンピュータ111にテキスト分類エンジンプログラムを記録した記録媒体112を備えるようにしたものである。その他の構成は、第2の実施の形態に係るアンケート回答分析システムと全く同様であるので、対応する部分には同一符号を付してそれらの詳しい説明を省略する。

[0106]

このように構成された第5の実施の形態に係るアンケート回答分析システムでは、記録媒体112からコンピュータ111にテキスト分類エンジンプログラムが読み込まれ、形態素解析手段15,カテゴリ・テキスト指定手段16,属性選択手段17,ルール学習手段18,およびルール出力手段19を含むテキスト分類エンジン83としてコンピュータ111の動作を制御する。コンピュータ111上でのテキスト分類エンジン83の動作の詳細は、第2の実施の形態に係るアンケート回答分析システムの場合と全く同様になるので、その詳しい説明を割愛する。

[0107]

(6)第6の実施の形態

図12は、本発明の第6の実施の形態に係るアンケート回答分析システムの全体構成を示すブロック図である。本実施の形態に係るアンケート回答分析システムは、図9に示した第3の実施の形態に係るアンケート回答分析システムにおいて、データベース93に接続されたコンピュータ121にテキスト分類エンジンプログラムを記録した記録媒体122を備えるようにしたものである。その他の構成は、第3の実施の形態に係るアンケート回答分析システムと全く同様であるので、対応する部分には同一符号を付してそれらの詳しい説明を省略する。

[0108]

このように構成された第6の実施の形態に係るアンケート回答分析システムでは

、記録媒体122からコンピュータ121にテキスト分類エンジンプログラムが 読み込まれ、形態素解析手段15,カテゴリ・テキスト指定手段16,属性選択 手段17,ルール学習手段18,およびルール出力手段19を含むテキスト分類 エンジン94としてコンピュータ121の動作を制御する。コンピュータ121 上でのテキスト分類エンジン94の動作の詳細は、第3の実施の形態に係るアン ケート回答分析システムの場合と全く同様になるので、その詳しい説明を割愛す る。

[0109]

【発明の効果】

[0110]

本発明の第1の効果は、企業に関するイメージ調査,特定の商品やサービス等の アンケートを請け負って、アンケートをネットワーク上で実施し、ネットワーク を通じてオンラインで集められた自然言語による自由回答記述を含むアンケート 回答文をデータベースに蓄え、そこからアンケート回答文を呼び出して、テキス ト分類エンジンを使用することにより得られるルール形式の知識を分析結果とし て要求者に販売できることである。

[0111]

本発明の第2の効果は、企業に関するイメージ調査,特定の商品やサービス等の アンケートを請け負って、アンケートを実施し、自然言語による自由回答記述を 含むアンケート回答文を一度に回収して、これをデータベースで蓄え、そこから アンケート回答文を呼び出して、テキスト分類エンジンを使用することにより得 られるルール形式の知識を分析結果として要求者に販売できることである。

[0112]

本発明の第3の効果は、企業に関するイメージ調査、特定の商品やサービス等の アンケートを請け負って、アンケートをネットワーク上で実施し、ネットワーク を通じてオンラインで集められた自然言語による自由回答記述を含むアンケート 回答文をデータベースで蓄え、そこからアンケート回答文を呼び出して、テキス ト分類エンジンを使用することにより得られるルール形式の知識を分析結果とし て、要求者からの要求に応じてネットワークを通じて配信サービスすることによ り販売できることである。

【図面の簡単な説明】

【図1】

本発明の第1の実施の形態に係るアンケート回答分析システムの構成を示すブロック図である。

【図2】

図1中のデータベースに蓄積されたアンケート回答文を例示する図である。

【図3】

図1中のテキスト分類エンジンにおける処理を示すフローチャートである。

【図4】

図3中の属性選択ステップのより詳細な処理を示すフローチャートである。

【図5】

図3中のルール学習ステップのより詳細な処理を示すフローチャートである。

【図6】

図1中のテキスト分類エンジンによる分析結果であるルール形式の知識 (確率的 決定リスト) の一例を示す図である。

【図7】

図1中のテキスト分類エンジンによる分析結果であるルール形式の知識(確率的 決定リスト)の他の例を示す図である。

【図8】

本発明の第2の実施の形態に係るアンケート回答分析システムの構成を示すブロック図である。

【図9】

本発明の第3の実施の形態に係るアンケート回答分析システムの構成を示すブロック図である。

【図10】

本発明の第4の実施の形態に係るアンケート回答分析システムの構成を示すブロック図である。

【図11】

本発明の第5の実施の形態に係るアンケート回答分析システムの構成を示すブロック図である。

【図12】

本発明の第6の実施の形態に係るアンケート回答分析システムの構成を示すブロック図である。

【符号の説明】

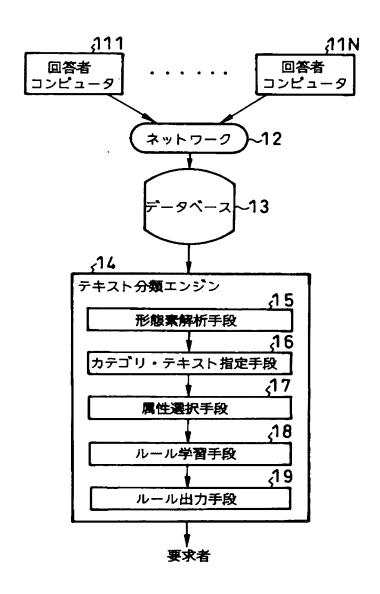
- 12 ネットワーク
- 13 データベース
- 14 テキスト分類エンジン
- 15 形態素解析手段
- 16 テキスト・カテゴリ指定手段
- 17 属性選択手段
- 18 ルール学習手段
- 19 ルール出力手段
- 31 形態素解析ステップ
- 32 テキストおよびカテゴリ指定ステップ
- 33 属性選択ステップ
- 34 ルール学習ステップ
- 35 ルール出力ステップ
- **41** ΔSC(ω)計算ステップ
- 42 属性選択ステップ
- 51 データ成形ステップ
- 52 成長処理ステップ
- 53 刈り込み処理ステップ
- 81 アンケート回答入力手段
- 82 データベース
- 83 テキスト分類エンジン
- 92 ネットワーク
- 93 データベース

特2000-07165

- 94 テキスト分類エンジン
- 95 要求者コンピュータ
- 101, 111, 121 コンピュータ
- 102, 112, 122 記録媒体
- 111~11N 回答者コンピュータ
- 911~91N 回答者コンピュータ

【書類名】 図面

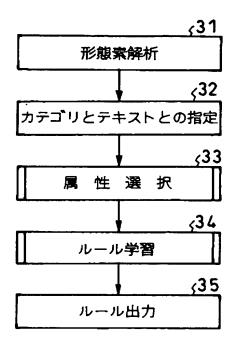
【図1】



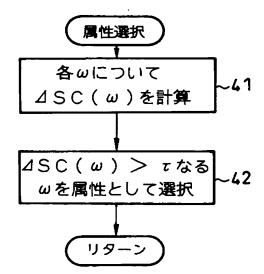
【図2】

アンケート 回答者	ハイテク企業として どこ を想起 しますか?	あなたにとってハイテク とは何ですか?	ハイテク商品として 何を思い浮かべますか?	
1	A社	進化した未来的な機械	ロボット	• • • • •
2	C社	使いやすく優しいもの	拼带電話	
3	A社	高速で高性能な機械	パソコン	
	•••			

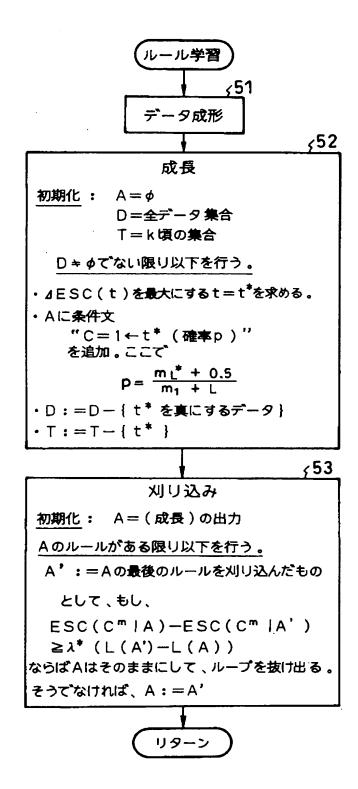
【図3】



【図4】



【図5】



【図6】

A社 ← 使いやすい (92.0%)

A社 ← 未来&プライベート(87.2%)

A社 ← 疲労&軽減〔78.0%〕

A社 ← 簡単だ (65.8%)

A社 ← 楽しい(56.2%)

A社でない ← それ以外 (79.4%)

【図7】

B社 ← 迅速だ (82.0%)

B社 ← 機械&効率〔77.8%〕

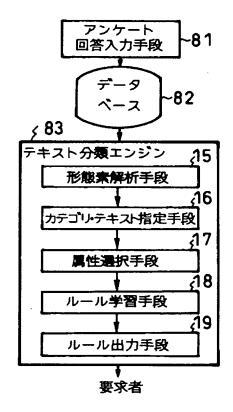
B社 ← 機械&操作〔76.0%〕

B社 ← 巧妙だ (60.8%)

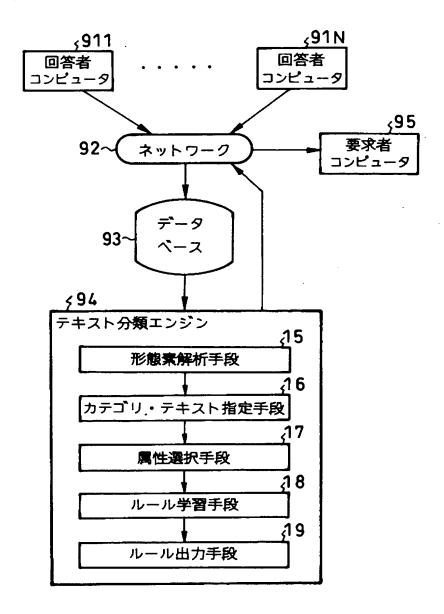
В社 ← 優れた〔60,2%〕

B社でない ← それ以外 (76.4%)

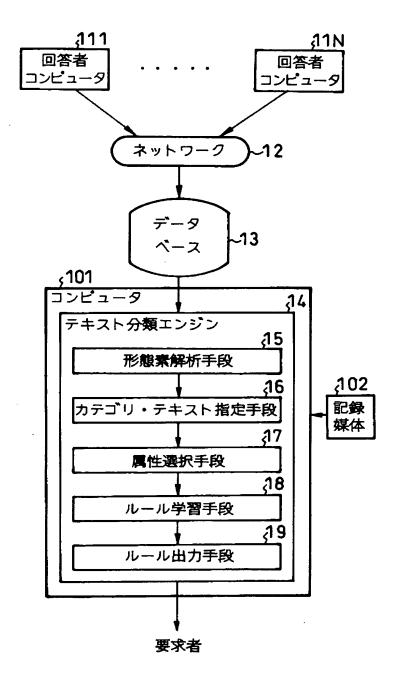
【図8】



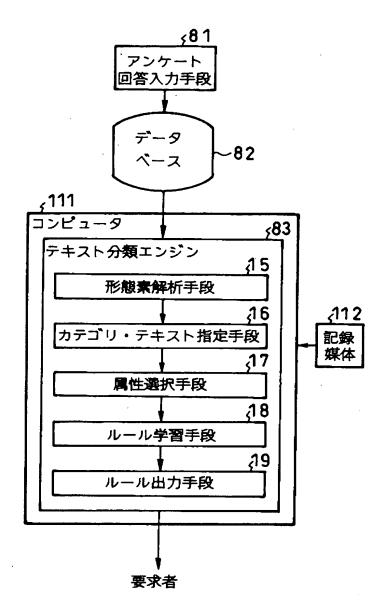
【図9】



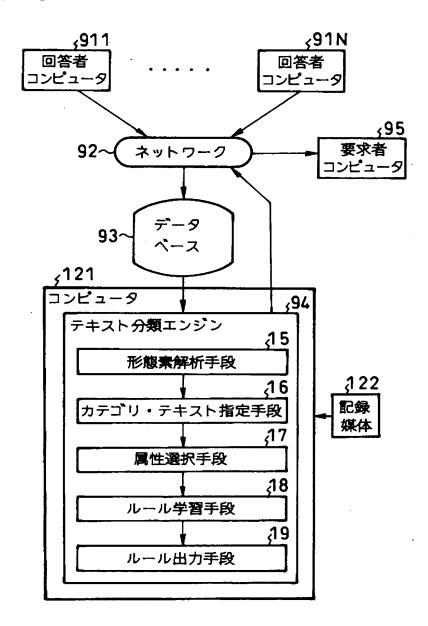
【図10】



【図11】



【図12】



【書類名】 要約書

【要約】

【課題】ネットワークを通じて回収した自然言語による自由回答記述を含むアンケート回答文から、テキスト分類エンジンを用いることにより、アンケート回答分析を自動的に行い、分析結果をルール形式の知識として要求者に出力する。

【解決手段】アンケート回答者は回答者コンピュータ 111~11Nからアンケート回答文を送信する。アンケート回答文は、ネットワーク12を通じてデータベース13に蓄積される。テキスト分類エンジン14は、データベース13から蓄積されたアンケート回答文を読み出して、アンケート回答文を分類するルールを学習して、要求者に出力する。

【選択図】 図1

認定・付加情報

特許出願の番号

特願2000-071657

受付番号

50000307659

書類名

特許願

担当官

第七担当上席

0096

作成日

平成12年 3月16日

<認定情報・付加情報>

【提出日】

平成12年 3月15日



出願人履歴情報

識別番号

[000004237]

1. 変更年月日

1990年 8月29日

[変更理由]

新規登録

住 所

東京都港区芝五丁目7番1号

氏 名

日本電気株式会社